Dynamic Spectrum Derived Mfcc and Hfcc Parameters and Human Robot Speech Interaction

Krishna Kumar Sharma[1], Piyush Kapoor[1], Prof. G C Nandi[1], and Dr Pavan Chakraborty[1]
[1] Indian Institute of Information Technology, Allahabad, India
Email: krisshna.sharma@gmail.com, piyushkapoor7@yahoo.com, gcnandi@iiita.ac.in, pavan@iiita.ac.in

***Abstract*—Using the Mel-frequency cepstral coefficients (MFCC), Human Factor cepstral coefficients (HFCC) and their new parameters derived from log dynamic spectrum and dynamic log spectrum, these features are widely used for speech recognition in various applications. But, speech recognition systems based on these features do not perform efficiently in the noisy conditions, mobile environment and for speech variation between users of different genders and ages. To maximize the recognition rate of speaker independent isolated word recognition system, we combine both of the above features and proposed a hybrid feature set of them. We tested the system for this hybrid feature vector and we gained results with accuracy of 86.17% in clean condition (closed window), 82.33% in class room open window environment, and 73.67% in outdoor with noisy environment.**

***Index Terms*— MFCC, HFCC, HMM, HOAP-2**

I. Introduction

In the present scenario, robotics are gaining an increasing role in the social life. Speech is a most natural way to communicate for human as compare to eye-gazing, facial expression, and gestures to interact with robot [1]. But speech recognition performance varies according to environment and users. Robots are mobile in nature and controlled by different users so, it should be noise robust, environment adaptability, and user adaptability for different ages and sex. To achieve noise robustness, environment adaptability, and user adaptability, features MFCC, ΔsMFCC,and ΔMFCC [2] are widely used as a feature vector for speech recognition. HFCC [3] is also has been used as a feature for speech recognition. HFCC outperform in the clean condition, but to make it efficient feature vector in the noisy condition we used dynamic cepstral coefficient of HFCC. As described above feature vector of MFCC are used for recognition and similarly HFCC are also separately used for recognition purpose, both features work smartly in different-different situations. MFCC filter bank and HFCC filter bank are different in design perspective, in MFCC filter bank spacing is dissociated with filter bandwidth but, in HFCC filter spacing is associated with equivalent rectangle bandwidth (ERB) that is introduced by Moore and Glasberg. Static MFCC and static HFCC features can attain high accuracy in the clean environment but, in case of robot environment, it is not always clean. It varies and has noise. So, to tune parameters with above mentioned conditions, dynamic parameters are used of cepstral coefficients. Dynamic MFCC i.e.ΔMFCC is the spectral filtered cepstral coefficient in the log spectral domain. And another feature is derived from log dynamic spectrum i.eΔsMFCC. And another updated feature is HFCC, and its dynamic HFCC i.e. ΔHFCC is the spectral filtered cepstral coefficient in the log spectral domain. And another feature is derived from log dynamic spectrum i.e.ΔsHFCC. Now, three set of feature vector are used to test recognition performance, those are respectively: 1) ΔMFCC + ΔsMFCC + MFCC, 2) ΔHFCC + ΔsHFCC + HFCC, and 3) ΔMFCC + ΔHFCC + MFCC + HFCC. Among these feature set number three performed best in recognition percentage with 86.17% in the lab environment (closed window), 82.33% in lab environment (open window) and 73.67% in outdoor noisy environment. Feature set number seven also performed good but here we have to filter data in two filter HFCC filter bank and MFCC filter bank, so it take more time to process data as compare to other feature set. After extracting features from the speech samples, we need to generate codebook from features. Linde–Buzo–Gray *algorithm* [4] is used to quantize features, this is a iterative technique of vector quantization. And then Hidden Markov Model (HMM) [5] technique is used to get good recognition result. To test the speech recognition system in the real time, a 25 degree of freedom humanoid robot HOAP-2 is used [6], This HOAP-2 is simulated in the WEBOTS real time simulation software [7]. The rest paper is organized as follows. In section 2, we describe speech recognition (SR) system, which contains techniques to extract different cepstral coefficients, vector quantization method, and HMM model design method. In section 3, we present proposed method. Section 4, describes results and comparison. And, Section 5 concludes the paper.

II. SR System

MFCC is widely used speech feature for automatic speech recognition. The functionality of MFCC is attributed to characteristics of the triangular sized filter bank as shown in Fig. 2. Calculated energy of each filter smoothes the speech spectrum, repressing the effects of pitch, and the warped frequency scale provides changeable sensitivity to the speech spectrum. But MFCC does not resemble the approximate critical bandwidth of human auditory system. HFCC is devised to dissociate filter bandwidth from number of filters and frequency range. When signal noise ratio is high, then MFCC and HFCC performed well, but in the noisy situation their performances degrade. To reduce the noise effect in the signal, we calculate their dynamic and static parameter of MFCC and HFCC. MFCC and HFCC are extracted as given in the Fig. 1. MFCC and HFCC differ only in the filter design technique. We used different combination of parameters. MFCC and its dynamic features and in the similar way HFCC and its dynamic features are used as shown in the Fig. 1. And combined features of MFCC and HFCC are used as a new

DOI: 01.IJEPE.02.02.69

proposed feature for this application as shown in the Fig. 1.

A. Feature Extarction

To calculate cepstral coefficients following steps are used as per shown in the Fig. 1. Let s(n) is the input speech signal recorded at 16KHz frequency with signed 16- bit resolution.

1) Signal s(n) is put through a low order digital system. to spectrally flatten the signal and to make less susceptible to finite precision effect later in the speech processing.

$$H(z) = 1 - a \times z^{-1}, \quad \text{where a=0.95} \qquad (1)$$

2) Speech signal is quasi-static signal, so it is divided in to frames of 25msec length for 16kHz speech signal [9].

In other words, we can also say that a frame contain FL=400 samples and next frame start followed by 160 samples or adjacent frames are being separated by 160 samples.

3) Now, framed signals are passed through hamming window to maintain continuity in the signal. There are other windows also but it efficiently remove side ripples in the signal.

$$Sw(n,\tau) = \{0.54 - 0.46 \times \cos(2\pi(n-1)/(FL-1))\} \times Sf(n,\tau). \qquad (2)$$

$1 \le n \le FL$, where τ is frame index, FL is the frame length.

4) Now, to get frequency spectrum of the signal, FFT is performed on the windowed signal.

$$S(K,\tau) = \left| \sum_{n=0}^{n=FL-1} Sw(n,\tau) \cdot e^{-jnK\frac{2\pi}{FL}} \right|,$$

$$K = 0,1 \ldots\ldots, FL-1 \qquad (3)$$

5) Hearing perception of human is not equally sensitive to all frequencies signal, it behave logarithmically to high frequencies signal and linearly to low frequencies signal. Filters are designed according to mel scale to the spectrum. Mel scale is approximately logarithmic for above 1kHz and linearly below 1kHz.mel frequency is calculated from following equation:

$$Mel(f) = 2595 \log_{10}(1 + f/700). \qquad (4)$$

$f(K)$ is denoted in the form of K, fs, and FL as follows:

$$f(K) = Kfs/FL, \qquad (5)$$

Here, FL is the frame length.

And, MFCC filter bank is designed as follows:

$$H(K,b) = \begin{cases} 0, & for\ f(K) < fc(B-1) \\ (f(K) - fc(b-1))/(fc(b) - fc(b-1)), & for\ fc(B-1) \le f(K) < fc(B) \\ (f(K) - fc(B+1))/(fc(B) - fc(B-1)), & for\ fc(B1) \le f(K) < fc(B+1) \\ 0, & for\ f(K) < fc(B-1) \end{cases} \qquad (6)$$

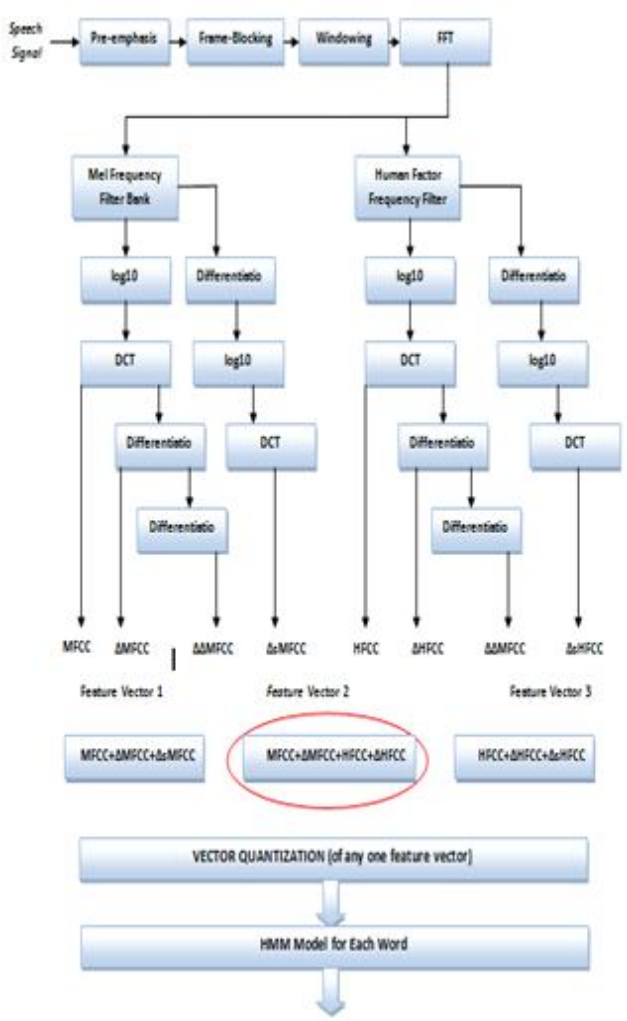

Figure. 1 Block Diagram from the feature extraction to design HMM model.

Here, $fc(b)$ is the center frequency of the filter and $H(K,b)$ is a group of triangular filters that have equal-height. Boundary points are uniformly spaced in the mel scale. MFCC filter bank is obtained as shown in the Fig. 2 for the 19 filters in the range of 8KHz.

6) The output of the mel filtering is passed to logarithmic function (natural logarithm) to obtain log-energy output.

$$Sm(B,\tau) = \ln\left(\sum_{K=0}^{K=FL-1} |S(K,\tau)| \cdot H(K,B)\right),\ B = 1,2,\ldots\ldots, M\ number\ of\ filter \qquad (7)$$

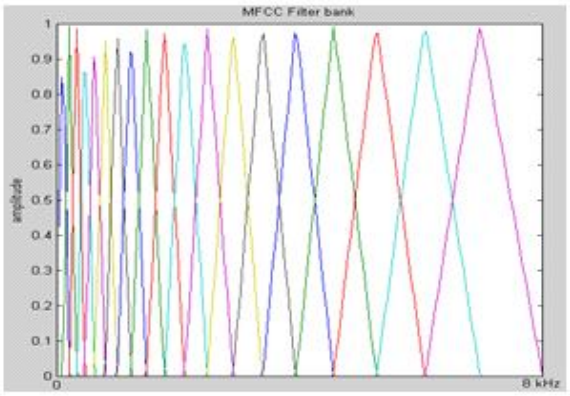

Figure 2. MFCC Filter bank of 19 filter in 8KHz range.

DOI: 01.IJEPE.02.02. 69

7) To obtain the static feature of MFCC Discrete Cosine Transform (DCT) is employed to log-energy.

$$c(j,\tau) = \sum_{B=1}^{B=M} Sm(B,\tau)\cdot\cos(j\cdot(B-1/2)\cdot\pi/M)\,,$$
$$j = 0,1,\dots\dots\dots,J. \qquad (8)$$

j is the index of the cepstral coefficient. M is the number of filter bank that are used and, J is the number of MFCC are used. This obtain MFCC are called static feature of speech, these are less immune to noise and changing environment condition. To make it more robust feature are extracted from dynamic spectrum as per shown in the figure 1.

8) HFCC features are also obtained by employing HFCC filter bank in place of MFCC filter bank in the feature extraction step 6. HFCC filter is designed as per described by *Mark D. Skowronski and John G. Harris* [9]. That will be static, dynamic log spectrum parameter and log dynamic spectrum parameter by applying further steps from 6 to 7.

9) After these steps, we got following features from the HFCC filter bank and MFCC filter bank:

MFCC, ΔMFCC, ΔsMFCC, ΔΔMFCC, HFCC, ΔHFCC, ΔsHFCC, ΔΔ HFCC. Now, we need to generate codebook for the different feature combinations as described in the INTRODUCTION part. codebook is generated from the . Linde–Buzo–Gray (LBG) vector quantization algorithm. 12 coefficients of each feature are used for each frame in the further processing, in the 6 type parameter set 3 features are used to make feature vector but, in one feature vector 4 features are used to make feature vector.

B. Vector Quantization

The LBG vector quantization is an iterative technique of quantization. The codebook is generated on the binary-splitting method, means initially average code vector is split in to two, and further in to 2^n vector, n is the splitting number.

Vector quantization is used to generate codebook for HMM. Following methodology is used to quantize vector:

1. Initially, calculate the centroid of each frame for the speech sample.
2. Now, split the centroid of codebook Yn according to:

$$Yn^{+} = Yn(1+\varepsilon)$$
$$Yn^{-} = Yn(1-\varepsilon) \qquad \text{Where } \varepsilon = 0.01 \qquad (9)$$

3. Find out the nearest – neighbor for each training vector. This is done using the K-mean iterative algorithm.
4. Update the centroid according to member of the cluster.
5. Repeat step 3 and 4 until the average distance falls below a preset threshold.
6. Repeat steps 2, 3, and 4 until codebook of size M is reached. M=2^n, n is the splitting number, and M is the desired size of the codebook.

C. Hidden Markov Model (Hmm)

After vector quantization of samples we get codebook to generate HMM model of words. Each word is spoken by 18 people in 10 different conditions and manner, so each word has 180 samples. Now, each word's 180 samples are vector quantized by LBG technique. From 180 samples we get 36 samples feature vector from LBG technique. And from these 36 samples we designed HMM model of each word. And finally, performing a viterbi search algorithm to find out most likely state sequence in HMM given a sequence of observed output.

III. PROPOSED METHOD

After successfully extracting features from MFCC coefficient and its dynamic coefficient are used as features. Similarly HFCC coefficient and its dynamic coefficient are also used as feature vector. As shown in the Fig. 1, the key difference between these two parameters is in the design of the filter bank that is described in the ASR system. To increase system efficiency we proposed possible combination of the features, which include both filter's characteristics, in the following steps:

Step 1: as described in the section 2.1, we obtained MFCC and HFCC features and their dynamic parameters also.

Step 2: according our proposed method, we make combined feature vector of both parameter and vector quantized the parameter.

Step 3: and from generated codebook after vector quantization, we develop HMM model of the each word.

Step 4: system is trained from step 3, now to use this system we find out testing samples maximum likelihood from the HMM models using viterbi algorithm.

HMM model θ is not similar to markov model whose states can not be directly observed [10]. HMM model can be described in to two sections; 1) one discusses about the entities associated with the HMM model and, 2)present technique of HMM based similarity measure.

HMM contain following entities:

1. $s = \{s1, s2, \dots\dots\dots\dots, sN\}$ is finite state sequence, where N is the number of states.
2. $A = \{aij\},\ 1 \le i,j \le N$, denoting the transition probability from si to sj. And $\sum_{j=1}^{N} aij = 1$. (10)
3. Emission matrix $B = \{b(O/S)\}$, denoting the probability of emission symbol O when state is S.
4. Initial state probability distribution

$$\pi = \{\pi i\},\ 1 \le i \le N \qquad (11)$$

denoting the probability of state Si.

HMM based similarity measure:

HMM θ is trained for each utterance O by baum-Welch algorithm [5]. HMM find out similarity between two utterances Oi and Oj by calculating similarity of θi and θj as:

$$D(\theta i,\theta j) = \frac{1}{2}\left(P(Oi/\theta j) + P(Oj/\theta i)\right) \qquad (12)$$

Where, $P(O/\theta)$ is calculated using the famous viterbi algorithm, in which log likelihood is used.

DOI: 01.IJEPE.02.02. 69

IV. EXPERIMENT AND RESULTS

The design of the Speech recognition system is varied according to their objective, like isolated speech recognition system, and continuous speech recognition system for two modes speaker dependent and speaker independent.

TABLE I.
FEATURE VECTOR 1

			Results from MFCC+ΔMFCC+ΔsMFCC		
S. No.	WORDS TO BE RECOGNIZED	MAX NUMBER OF ATTEMPS (BY DIFFERENT USERS)	RCOGNIZED RESULT (IN THE CLOSED WINDOW LAB)	RCOGNIZED RESULT (IN THE OPEN WINDOW LAB)	RCOGNIZED RESULT (IN THE OUTDOOR NOISY ENVIRONMENT)
1	START	40	34	31	29
2	STOP	40	35	32	28
3	FORWARD	40	34	33	27
4	BACKWORD	40	34	33	29
5	LEFT	40	36	32	30
6	RIGHT	40	34	32	29
7	UP	40	33	33	29
8	DOWN	40	34	33	27
9	NAMSTE	40	33	32	28
10	BYE-BYE	40	34	31	26
11	ONE	40	34	32	27
12	TWO	40	33	33	26
13	THREE	40	34	32	28
14	FOUR	40	35	30	26
15	FIVE	40	35	33	27

For the training of the isolated speech recognition system of the speaker independent mode, we gathered speech samples of the different age people, different environment (lab room with close window, lab room with open window or class room environment, and outdoor noisy environment) [11]. To test this recognition system, we used humanoid robot HOAP-2. In the real time we commanded humanoid robot and observed results. We collected each word spoken by 18 different users 10 times by each user or in other words 180 times each word. We categorized users in to the three age categories 6-13 years, 14-50 years, and 51-70 years of both genders. We processed the collected data and trained the system and tested it for the other speaker other than the trained sample's speakers.

TABLE II.
FEATURE VECTOR 3

			Results from HFCC+ΔHFCC+ΔsHFCC		
S. No.	WORDS TO BE RECOGNIZED	MAX NUMBER OF ATTEMPS (BY DIFFERENT USERS)	RCOGNIZED RESULT (IN THE CLOSED WINDOW LAB)	RCOGNIZED RESULT (IN THE OPEN WINDOW LAB)	RCOGNIZED RESULT (IN THE OUTDOOR NOISY ENVIRONMENT)
1	START	40	34	32	27
2	STOP	40	34	31	28
3	FORWARD	40	36	34	27
4	BACKWORD	40	36	34	27
5	LEFT	40	35	32	31
6	RIGHT	40	33	31	30
7	UP	40	34	32	28
8	DOWN	40	34	33	28
9	NAMSTE	40	35	34	31
10	BYE-BYE	40	33	33	27
11	ONE	40	33	31	26
12	TWO	40	34	33	29
13	THREE	40	34	32	27
14	FOUR	40	34	34	26
15	FIVE	40	35	32	29

Form this experiments we get good results for our own dataset that is described in the following tables Table 1, Table 2, and Table 3. In the Table 1, we described about result obtained from the MFCC, ΔsMFCC, and ΔMFCC features. In the Table 2, we described about result obtained from the HFCC, ΔsHFCC, anc ΔHFCC features. And In the Table 3, we described about result obtained from the MFCC, HFCC, ΔMFCC, anc ΔHFCC features. From the obtained result we found that in the clean environment of the lab in the closed window form then recognition rate is less varied for all three feature vector, that are respectively: 85.33%, 85.67%, and 86.17% for Table 1, Table 2 and Table 3. And if we increase the noise ratio in the signal than this recognition rate change slightly. Now, we evaluated in the class room noisy or open window form, obtained results are as followed 80.33%, 81.33%, and 82.33% from three tables respectively Table 1, Table 2, and Table 3. And we also tested for the real time environment in the outdoor noisy environment for the all feature sets, obtained results are as followed 69.33%, 70.17%, and 73.67% from Table 1, Table 2, and Table 3. From obtained results, we can find out that feature set of combined filter work better as compare to the single filtered feature sets. This is like features are facilitated with the characteristics of the both filters and they efficiently resemble human auditory system because MFCC filter bank works linearly for below 1 KHz signal and logarithmic for above 1 KHz signal. And HFCC filter bank also resemble human auditory system with modified filter bank spacing than MFCC.

TABLE III.
FEATURE VECTOR 2

			Results from MFCC+ΔMFCC+HFCC+ΔHFCC		
S. No.	WORDS TO BE RECOGNIZED	MAX NUMBER OF ATTEMPS (BY DIFFERENT USERS)	RCOGNIZED RESULT (IN THE CLOSED WINDOW LAB)	RECOGNIZED RESULT (IN THE OPEN WINDOW LAB)	RECOGNIZED RESULT (IN THE OUTDOOR NOISY ENVIRONMENT)
1	START	40	34	32	30
2	STOP	40	35	34	30
3	FORWARD	40	34	33	28
4	BACKWORD	40	34	33	29
5	LEFT	40	35	33	30
6	RIGHT	40	36	35	31
7	UP	40	33	33	29
8	DOWN	40	34	33	29
9	NAMSTE	40	33	32	30
10	BYE-BYE	40	34	32	28
11	ONE	40	35	33	30
12	TWO	40	34	33	30
13	THREE	40	34	32	29
14	FOUR	40	35	33	30
15	FIVE	40	37	33	29

V. CONCLUSION AND FUTURE WORK

In This paper, we described about different features based speaker independent isolated speech recognition system, and find out there efficiency in the different mobile environment and user adaptability. Features extraction only differ in the filtering process based on their filter bank construction, and we find out that efficiency of recognition increase as we combine different features and their differentiated parameters

DOI: 01.IJEPE.02.02.69

as compare to one filtered parameter. Combined features set of MFCC and HFCC take more computing time in the training phase and testing as compare to individual features set of MFCC and also from HFCC features set. But, results improved in the combined feature set as shown in the Table 1, Table 2, and Table 3. In the future work, we would like to use this technique for large number of words, the continuous speech recognition, and would like to make it more robust for the different races people also.

REFERENCES

[1] Carlos Toshinori Ishi, Shigeki Matsuda, Takayuki Kanda, Takatoshi Jitsuhiro, Hiroshi Ishiguro, Satoshi Nakamura, and Norihiro Hagita, "A Robust Speech Recognition System for Communication Robots in Noisy Environments," IEEE Transactions on robotics, vol. 24, no. 3, June 2008.

[2] Nengheng Zheng, Xia Li, Houwei Cao, Tan Lee, and P. C. Ching,"Deriving MFCC parameters from the dynamic spectrum for robust speech recognition", Chinese Spoken Language Processing, 2008. ISCSLP '08. 6th International Symposium on.

[3] Mark D. Skowronski and John G. Harris, "Improving the filter bank of a classic speech feature extraction algorithm," IEEE Intl Symposium on Circuits and Systems, Bangkok, Thailand, vol IV, pp 281-284, May 25 - 28, 2003.

[4] WEB,"http://www.softwarepractice.org/wiki/Vector_Quantization_LBG_Pseudo_Code"

[5] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition.Proc. of IEEE, 77(2):257-286,1989.

[6] WEB, "http://robita.iiita.ac.in/hoap2instruction03e.pdf"

[7] WEB, "http://www.cyberbotics.com/"

[8] Iosif Mporas, Todor Ganchev, Mihalis Siafarikas, Nikos Fakotakis, "Comparison of Speech Features on the Speech Recognition Task," Journal of Computer Science 3 (8): 608-616, 2007.

[9] ETSI ES 201 108, V1.1.3 (2003-09). ETSI standard: speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms; (Sect. 4, pp. 8–12).

[10] M.Gales and SJ Young.The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing- 1(3), pp. 195-304,2008.

[11] Hong Liu, Xiaofei Li, "A Selection Method of Speech Vocabulary for Human-Robot Speech Interaction" Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on , vol., no., pp.2243-2248, 10-13 Oct. 2010

DOI: 01.IJEPE.02.02. 69